Data Analytics Assignment 2 Report

Dadi Sasank Kumar, 22CS10020
Sumit Kumar, 22CS30056
Vangapandu Tejaraam, 22CS30059

1 Introduction

This report details the implementation and results of association rule mining on a grocery dataset using the Apriori algorithm. The goal is to identify frequent itemsets and generate association rules that can provide insights into customer purchasing behavior.

2 Dataset

The dataset used for this assignment is a grocery dataset containing transactions of items purchased. The dataset was loaded into a pandas DataFrame and preprocessed to create a list of transactions.

3 Preprocessing

The dataset was preprocessed to handle missing values and convert the transactions into a suitable format for the Apriori algorithm. The preprocessing steps included:

- Loading the dataset into a DataFrame.
- Handling missing values.
- Creating a list of transactions where each transaction is a set of items.

4 Unique Items

We identified the unique items present in the transactions and calculated their count.

```
unique_items = get_unique_items(transactions)
print(unique_items)  # Output: {1, 2, 3, 4, 5}
print("Number of unique items:", len(unique_items))
```

Unique Items:

Unique Items(Output)

Number of Unique Items: 152

5 Apriori Algorithm

The Apriori algorithm was implemented to generate frequent itemsets. The algorithm was enhanced with dynamic minimum support thresholds based on item categories.

5.1 Dynamic Minimum Support

A category support map was used to apply dynamic minimum support thresholds for different categories of items. This approach helps in identifying frequent itemsets more effectively by considering the nature of the items.

5.2 Frequent Itemsets

The frequent itemsets were generated using the Apriori algorithm with the specified minimum support threshold.

```
min_support = 0.3
frequent_itemsets = apriori(transactions, min_support, category_support_map=category_support_map)
```

Frequent Itemsets and Support Values(Output)

6 Association Rules

Association rules were generated from the frequent itemsets using a minimum confidence threshold. The rules were enhanced with profitability weighting to prioritize more profitable items.

Association Rules

6.1 Profitability Mapping

A profitability map was used to assign profitability values to different items.

```
profitability_map = {
    # High Profit
    'specialty cheese': 0.09, 'red/blush wine': 0.09, 'white wine': 0.09, 'sparkling wine': 0.09,
    'liquor': 0.09, 'prosecco': 0.09, 'brandy': 0.09, 'rum': 0.09, 'specialty chocolate': 0.09,

    # Moderate Profit
    'whole milk': 0.06, 'yogurt': 0.06, 'fruit/vegetable juice': 0.06, 'bottled water': 0.06,
    'domestic eggs': 0.06, 'root vegetables': 0.06, 'detergent': 0.06, 'coffee': 0.06,
    'canned fish': 0.06, 'tea': 0.06, 'beef': 0.06, 'pork': 0.06,

    # Low Profit
    'rolls/buns': 0.03, 'brown bread': 0.03, 'white bread': 0.03, 'potato products': 0.03,
    'canned beer': 0.03, 'napkins': 0.03, 'cling film/bags': 0.03, 'bathroom cleaner': 0.03,
    'canned vegetables': 0.03, 'salt': 0.03, 'margarine': 0.03, 'soda': 0.03,

    # Promotion-Focused
    'snack products': 0.05, 'candy': 0.05, 'frozen dessert': 0.05, 'pastry': 0.05,
    'ice cream': 0.05, 'popcorn': 0.05, 'specialty bar': 0.05, 'salty snack': 0.05,
    'chocolate': 0.05, 'soft drinks': 0.05,

    # General Groceries (defaults to mid-range)
    'processed cheese': 0.05, 'cream cheese': 0.05, 'jam': 0.05, 'butter': 0.05,
    'rice': 0.05, 'sausage': 0.05,

    # Catch-all for items not specifically categorized
    'default': 0.04
}
```

Figure 1: Profitability Map

6.2 Generated Rules

The rules were generated with the specified minimum confidence threshold and profitability weighting.

```
min_confidence = float(input("Enter minimum confidence threshold (as a decimal): "))
rules_df = generate_rules(frequent_itemsets, min_confidence, transactions, profitability_map)
```

***Generated Association Rules*(from terminal)**

7 Visualization

7.1 Top-N Frequent Itemsets

The top-N frequent itemsets were visualized using a bar chart.

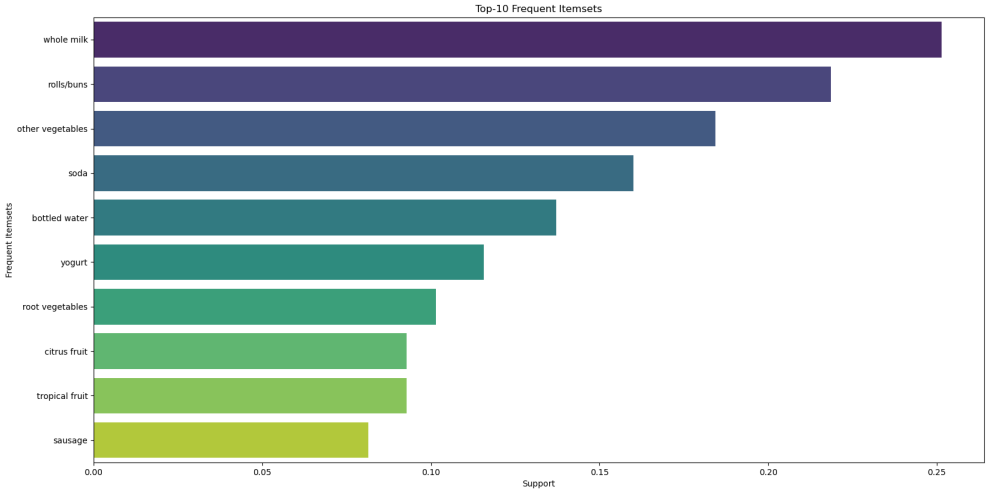

Figure 2: Top-N Frequent Itemsets

7.2 Top-N Strongest Rules

The top-N strongest rules were visualized using a scatter plot, showing lift and leverage values.

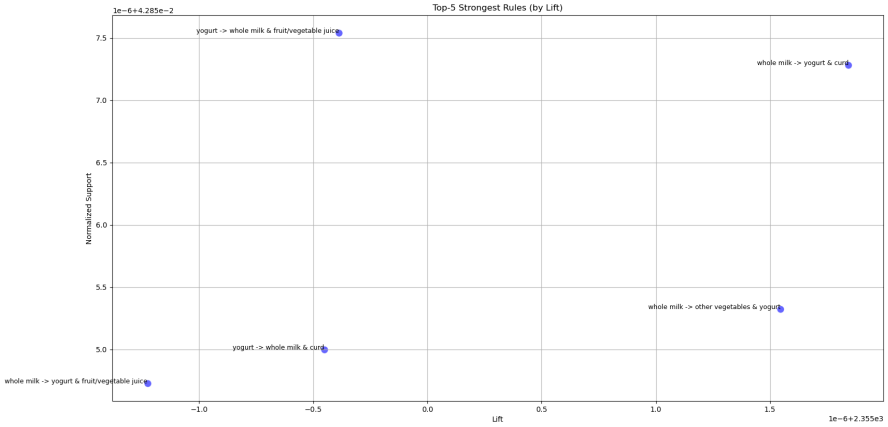

Figure 3: Top N Strongest Rules

8 Summary

This report presents the results of association rule mining on a grocery dataset using dynamic minimum support thresholds and profitability weighting to enhance the Apriori algorithm. The generated rules and visualizations offer insights into customer purchasing patterns, which can assist in targeted marketing and inventory management.

9 Advanced Modifications

- **Dynamic Minimum Support:** Different thresholds for various item categories enhance the identification of frequent itemsets.
- **Profitability Weighting:** Prioritizing profitable items ensures that generated rules are both frequent and valuable from a business standpoint.